I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

P2P on Campus: Who, What, and How Much

ALEXANDRE M. MATEUS* AND JON M. PEHA**

Abstract: This Article presents findings from a large-scale quantitative assessment of online exchanges of copyrighted material on a college campus based on network data collected using deep packet inspection. Findings show that use of Peer-to-Peer (P2P) to transfer copyrighted content is widespread on campus, although observed P2P is declining. In a month-long period in Spring 2008, 40% of students living on campus were observed engaging in P2P and 70% of those were detected attempting to transfer copyrighted content. Compared to Spring 2007, there was a 10% decrease in the daily percentage of detected P2P users and a 20% decrease in the daily percentage of P2P users detected attempting to transfer copyrighted content. P2P activity was detected across all genders, ages, classes and majors, and the observed decline was fairly uniform across demographics, except that P2P usage is high among freshmen in their first month on campus, and falls in subsequent semesters, indicating students became P2P users prior to college. The effect of this widespread P2P usage on content sales is complex. While 22% of P2P users purchased content from the iTunes Store (iTS), each buying on average about as much content as iTS customers who

* Alexandre M. Mateus, PhD in Engineering and Public Policy, Carnegie Mellon University, U.S.A., and Instituto Superior Técnico, Portugal, amfmateus@gmail.com, http://andrew.cmu.edu/~amateus.

** Jon M. Peha, Carnegie Mellon University, Professor in the Department of Engineering & Public Policy and the Department of Electrical & Computer Engineering, www.ece.cmu.edu/~peha. Dr. Peha contributed to this work in his capacity as a professor at Carnegie Mellon University, and dissertation advisor to Alexandre Mateus. Any opinion expressed herein is that of one or both authors, and does not represent the views of Carnegie Mellon University, the Federal Communications Commission, or any other organization.

didn't use P2P, 10% of P2P users only sampled content from the iTS, but did not buy. The fact that many students use both P2P and iTS shows that factors other than price and fear of lawsuits come to play in users' decisions, and content sellers might take advantage of such factors to win back some of their potential customers from P2P. As for the content itself, P2P users transferred a diverse mix including music, video and software. Music and video transferred over P2P show a heavy-tailed distribution of popularity, with less popular content accounting for a significant share of transfers. Moreover, for video, the importance of less popular content is growing. This indicates that a legal seller may need a large inventory to compete with P2P.

I. INTRODUCTION

Peer-to-Peer (P2P) networks are used to illegally transfer copyrighted content, although opinions differ as to how large the phenomenon is, how it is evolving, how it affects copyright holders, and how to deal with it. Copyright holders state that P2P piracy heavily impacts their revenues (RIAA, 2007). They have taken legal action against P2P developers and file sharers in the past, and continue to lobby for legislation against file sharing. College students are among the biggest users of file sharing (Lamy et al., 2007), thus drawing attention to P2P in university campuses. Some university networks in the United States have recently assumed a larger role in combating illegal transfers of copyrighted media over P2P by deploying network monitoring technology in an attempt to detect illegal transfers. Such an approach to deal with unauthorized exchange of copyrighted content online has been mandated for Internet service providers (ISPs) in a few countries (Bangeman, 2008a; Bremner, 2008), and similar mandates have been proposed in the United States and elsewhere for ISPs and universities (Bangeman, 2007; Fischer, 2007).

This Article focuses on online exchanges of media in university campuses, with an emphasis on exchanges using P2P and the iTunes Store. It seeks to fulfill three main objectives. First, to quantify the extent of P2P usage and transfers of copyrighted content using P2P on campus, how these are changing over time, and how they vary by the demographics of users, to help assess the need for intervention. Second, to characterize the content that students are obtaining from P2P, both to better understand the evolution of the P2P "market" for digital content, which can inform content providers of what they are competing against, as well as to identify possible drivers for P2P usage. Finally, this Article aims to shed light on the impact of P2P

transfers on revenues of copyright holders, which depends on the extent to which those transfers displace sales of content. This is done by quantifying the extent to which media is obtained from P2P and from the iTunes Store and by correlating usage of these sources.

The extent of P2P and illegal transfers of copyrighted content using P2P on university campuses has been previously assessed. By means of surveys studies found that over half of college students engaged in P2P file sharing (Lamy et al., 2007) and that college students amounted to 21% all P2P users (Guess, 2008; Oster, 2008). Moreover, these studies revealed that a significant share of students' media libraries were composed of music obtained from P2P file sharing (U.S. Congress, 2007a, testimony of Cary Sherman; Smith et al., 2007) and that college students obtained more of their music from P2P than the rest of the population (Lamy et al., 2008). Results from such survey-based studies depend on the memories and openness of survey respondents, how survey instruments are designed, and how the subjects are selected. This is particularly relevant in the task at hand, given that the subject in question constitutes illegal activity, and some respondents may refrain from disclosing their behavior in surveys. This Article presents results from a quantitative assessment of online media transfers based on actual observation of P2P exchanges on a college campus. Thus, not only are the results herein independent of whether or not survey respondents fully disclose their behavior, it is also possible to access information that Internet users may not know, such as the volume of P2P transfers or the time of such transfers.

The remainder of this Article is organized as follows. Section II presents relevant background to frame the analysis. Section III presents an overview of the data collection methodology, followed by a summary of the collected data in Section IV. Section V is the main results section, which first draws the general picture of P2P usage and how it is evolving on campus, followed by a demographic breakdown of P2P figures. Then it focuses on the content being transferred using P2P, and finally on how usage of P2P to obtain media compares to usage of the iTunes Store. This Article concludes with a summary of findings and policy implications in Section VI.

II. BACKGROUND

By U.S. Copyright, except where "fair use" provisions apply,[1] transfers of copyright-protected works without permission from the

[1] Copyright Act, 17 U.S.C. § 107.

copyright holder are infringements of the holder's rights.[2] Both those who transfer the copyright-protected works and those who aid and support such transfers can be held liable for copyright infringement. This means that both P2P users and P2P developers may be liable for copyright infringement. The Digital Millennium Copyright Act (DMCA) has provisions limiting Internet Service Provider (IPSs) liability under certain circumstances; but to obtain such "safe harbor" protection, ISPs must respond to subpoenas and identify subscribers accused of a violation.[3] In the past, the music industry, through the Record Industry Association of America (RIAA), used these legal provisions in several lawsuits against P2P companies (Macavinta, 1999; Oswald, 2006) and users (EFF, 2008). To unveil the identity of users, RIAA traditionally used the subpoena mechanism in the DMCA. "Since early 2007, the music industry has utilized "pre-litigation settlement letters" requesting that universities identify infringing student-users." (Buskirk, 2007). Since these letters were not legally binding, some universities ignored them, while others forwarded them to students (EFF, 2008). Upon receipt of the letters, students could avoid court action and settle the case using the phone or a website.[4] More recently, the industry reported the end of their lawsuit campaign against P2P users (McBride and Smith, 2008), and announced plans to start collaborating with ISPs for copyright protection.

To justify legal action against P2P users, music industry representatives assert that P2P is responsible for billions of dollars in lost sales and thousands of lost jobs.[5] There is a growing body of literature attempting to assess whether file sharing does indeed lead to a decline in sales of content, particularly in sales of music and video. However, as summarized in a recent working article (Oberholzer-Gee and Strumpf, 2009), different authors present contradictory results. Most articles focusing on this subject argue that P2P file sharing contributes to the decrease in music and video sales,[6] with displacement rates ranging between 3.5% (for movies) and 30% (for music), i.e., each music title downloaded through P2P displaces

---

[2] 17 U.S.C. § 501.

[3] 17 U.S.C. § 512.

[4] https://www.p2plawsuits.com

[5] U.S. Congress, 2007a, testimony of Cary Sherman.

[6] *See, e.g.*, Hong 2007; Liebowitz 2008; Michel 2006; Rob and Waldfogel 2006; Zentner 2006.

sales of 30% of a music title. But other articles argue that P2P has positive effects on music sales (Andersen and Frenz, 2008; Gopal and Bhattacharjee, 2006). Accounts of P2P file sharing in these articles are based on self-reported data collected by means of surveys or on use of proxies such as Internet penetration. A third set of authors, who use actual measures of file sharing, argue that P2P transfers are unrelated to changes in content sales (Bhattacharjee et al., 2007; Oberholzer-Gee and Strumpf, 2007; Smith and Telang, 2008; Tanaka, 2004). This article, through observation of online student usage of P2P as well as student usage of the iTunes Store, seeks to contribute some empirical evidence about the extent to which P2P users still purchase content online.

P2P is a rather promising technology for the distribution of content online because it allows the burden of distributing the content to be shared among the users who are obtaining such content. There are many legal uses for P2P, some of the higher profile ones being distribution of new releases of the Linux Operating System, distribution of patches and updates to popular games (e.g., World of Warcraft), or actual distribution of video by companies such as Vuze. But P2P continues to be widely used for the unauthorized transfer of copyrighted content. Part of the analysis in this article attempts to unveil whether there are motivations for use of P2P on campus other than for unlawfully exchanging copyrighted content.

Illegal file sharing has been debated at the legislative level in the U.S., Europe, and elsewhere. In the U.S., Congress held at least six hearings on online copyright infringement in universities between 2003 and 2007[7] and discussed possible interventions to deal with it.[8] Outside of the U.S., the focus was mostly on ISPs, particularly in the European Union where France approved legislation requiring ISPs to disconnect users detected transferring copyrighted material (Anderson, 2009a, 2009b; Pfanner, 2009), and the United Kingdom is discussing the possibility of similar legislation (Anderson, 2008; Wray, 2009). All this activity, both in the U.S. Congress and in some of the larger European Union countries, indicates potential for policy change. The characterization of P2P activity in a university campus provided in this article can contribute to and enlighten the discussion of new policies in this area.

---

[7] U.S. Congress 2003, 2004, 2005, 2006; 2007a, b.

[8] Bangeman, 2007; Fischer, 2007.

III. METHODOLOGY

This research was performed on data collected in the scope of the Digital Citizen Project (DCP), a project undertaken by Illinois State University (ISU) "to significantly impact illegal piracy of electronically received materials, using a comprehensive approach to confront pervasive attitudes and behaviors in peer-to-peer downloading of movies, music, and media."[9] In February 2007, a team of engineers and social scientists from Carnegie Mellon University (CMU) began conducting research on the dissemination of copyrighted material on the ISU campus. This section briefly describes the network monitoring procedures used to collect the data analyzed in this article.[10]

The ISU network serves the entire campus population. Students that live on campus purchase network access from ISU, which allows one wired connection per user in the dorm room to which a fixed IP address is assigned for the entire semester. Wireless routers are not allowed. Connectivity to the outside Internet is provided by two commercial Internet Service Providers (ISPs) and Internet2.[11] ISU uses traffic shaping in the connection to the ISPs and does not impose limits on the amount of traffic generated by each network user.

The network was monitored using two commercially available monitoring appliances that use deep packet inspection (DPI) to log relevant attributes of the communication between users inside the campus network and outside parties for traffic routed using commercial ISPs. Further monitoring was conducted inside the campus network using NetFlow collectors, whose output was analyzed using CMU-developed software. Such analysis is beyond the scope of the current article. One of the DPI appliances used for monitoring, Packeteer PacketShaper[12] (Packeteer), had already been deployed before this project to perform traffic shaping. It classifies

---

[9] Illinois State Uniersity, 2006.

[10] For a more detailed technical description of the data collection procedures, please refer to Mateus and Peha, "Characterizing digital media exchanges in a university campus network" (2009).

[11] Internet2 is an advanced not-for-profit networking consortium led by members from the research and education communities, industry, and government.

[12] Packeteer was since acquired by Blue Coat. For more information on the features of Packeteer PacketShapper (now Blue Coat Packetshapper), refer to http://www.bluecoat.com/products/packetshaper.

communication sessions in over 500 classes[13] according to the type of traffic that composes them without retaining the actual contents of the communications sessions. The other appliance, Audible Magic CopySense[14] (AM), was purchased to enforce ISU policy before CMU became involved. It uses header information to identify P2P streams, within which it tries to identify copyrighted media in real time. Such identification works by matching the transferred material against a database of audio fingerprints of copyrighted media titles[15] or hash codes[16] used to identify files in P2P networks. The AM device does not retain any portion of the transmission, but it does record which copyrighted material was matched in the database. When the material being transferred cannot be matched against anything in the database, AM records a piece of the metadata incorporated in the transfer (typically the name of the file being transferred).

Users and devices responsible for online activity detected by AM were identified using data from several network management and directory databases available only to ISU. Several measures were taken to protect the privacy of users in accordance with the DCP policy guidelines, and both the ISU Institutional Review Board (IRB) and the CMU IRB approved the research described in this article. The following measures were among those taken to protect the privacy of students. Data collection and anonymization were performed at ISU by the network management team, which already had access to the

---

[13] Each communication session originates one collected data record, which is a Netflow v.5 record augmented with class identifiers of the protocol/application used in the communication. For a list of fields typically contained in a Netflow v.5 records consult http://is.gd/gQuLx. Protocol/application classes include, among others, common protocols, services, Peer-to-Peer networks and content distribution networks. A detailed list of the classes available in the Packeteer version used for data collection can be found in https://bto.bluecoat.com/packetguide/7.3/reference/services.htm.

[14] For more information on the features of AM CopySense, refer to http://is.gd/gQuOx.

[15] One technique used by AM to identify copyrighted material is audio fingerprinting. AM collects a sample of the audio track of the material that is being transferred and extracts relevant and unique characteristics of that audio track (which are format- and encoding quality-independent). These are then compared against the database with the audio characteristics of known copyrighted titles.

[16] In most P2P networks, each file that is shared is identified using a unique hash code calculated based on the contents of the file. This short code (128 or 256 bytes) guarantees that the same file (i.e., the same content) is identified in the network independently of different filenames that it may have. The hash code is used by AM to identify copyrighted material because it allows for faster comparisons and earlier detection than the technique based on audio fingerprinting.

raw data. Both processes were automated and ran without human intervention. ISU staff that oversaw data collection were precluded from analyzing the anonymized data, and the only output provided to researchers at CMU was an anonymized version of the collected data. The anonymization process consisted of removing personally identifiable fields from the data to make it impossible to unveil the identity of ISU network users. In order to assign detected activity to single users or single computers, student identifiers, IP addresses and MAC addresses were replaced by pseudonyms generated using a one-way 256-bit hashing function,[17] whose hashing keys were destroyed. The only demographic characteristics of students that were collected were birth year, class, gender, and major. CMU researchers who analyzed the resulting data were not allowed to observe raw data prior to anonymization, and therefore were unable to connect any of the data to a specific person, computer, or location on campus.

IV. SUMMARY OF COLLECTED DATA AND DEFINITIONS

AM and Packeteer collected data during three periods of about one month each in the Spring 2007, Fall 2007, and Spring 2008 academic terms. In each of the periods, AM collected a log of events identified as being P2P traffic. In Spring 2007, Packeteer collected hourly summaries detailing the total amount of bytes and communication sessions entering and exiting the ISU network, broken down by protocol/application. In Fall 2007 and Spring 2008, Packeteer collected one individual record per detected communication session. Table 1 summarizes the data collected by each appliance in each period.

[17] Function $F(K,X) \rightarrow Y$ that, given a key K and an argument X, generates Y, a 256-bit long representation of X. F minimizes the probability that different X arguments will return the same Y. Furthermore, it is, in practical terms, impossible to map back from Y to X.

Table 1

Summary of data collected in the three monitoring periods by AM and Packeteer.

	Spring 2007	Fall 2007	Spring 2008
Number of people living on campus	6,544	6,764	6,763
Time span of AM data	Mar 1 to Apr 30	Sep 1 to Oct 4	Feb 12 to Apr 27
Full hours/days with AM data	648/25	654/26	1,747/60
Number of am events collected	24.6 million	22.2 million	58.1 million
Time span of Packeteer data	Apr 01 to Apr 30	Aug 30 to Oct 1	Mar 7 to May 1
Full hours/days with Packeteer data	720/30	735/29	858/31
Number of Packeteer events collected	Hourly summaries	3.3 billion	4.3 billion
Hours/days with both AM and Packeteer data	642/25	541/20	770/24

Data collected through network monitoring is always dependent on the monitoring devices visibility of the network being monitored. In this case, both monitoring appliances were deployed at the point where the campus network connects to commercial ISPs, which means that only communication sessions in which one party is inside the campus network and another party is in the external Internet can be detected. Hence, none of the data collected by AM or Packeteer include intra-campus communication sessions, nor communications routed through Internet 2.

The analysis in this article concerns two main types of activity: the usage of P2P and the usage of P2P to transfer copyright-protected media. A P2P activity is a communication event detected by AM or Packeteer in which information is transferred using a P2P protocol. A P2P user is a network user detected doing at least one P2P activity in one monitored period. A Detected Attempt to Transfer Copyrighted Media (DATCoM) is a detected AM event corresponding to a transfer or attempted transfer, using a P2P protocol, of media identified as being protected by copyright. A DATCoM user is a user who is detected doing at least one DATCoM in one monitored period.

Not every DATCoM is a copyright violation as defined by U.S. copyright law.[18] For instance, in some DATCoMs, users may be downloading material that will be used in particular ways that fall

[18] *See* 17 U.S.C. § 501.

under the "fair use" doctrine.[19] It is impossible to tell whether the copyrighted material in each DATCoM was used in any way that can be considered "fair use;" therefore, such considerations are outside the scope of this article. Also, the fact that detection may occur by matching the hash code in the P2P request to a database of hash codes of copyrighted content allows for the existence of some DATCoMs that correspond to P2P requests that never received a reply, in which no actual copyrighted content was transferred. However, for such a request to exist, one of the parties had to advertise that she was making copyrighted content available[20] and the other party had to search for that content and instruct her P2P client to download it. Results herein do not change significantly if such "empty" DATCoMs are disregarded because nearly all DATCoM users and copyrighted titles were detected in multiple DATCoMs. At least one of these DATComs contained enough bytes to actually correspond to a transfer, as opposed to a failed request. Hence, while not all DATCoMs detected on campus are copyright violations, most of them probably are, and they are good indicators that users engaged in illegal transfers of copyrighted content using P2P networks.

A DATCoM represents an attempt to transfer content without distinguishing downloads from uploads. There is no distinction between uploads and downloads because activities detected by AM do not contain conclusive information about direction of transfers. In legal terms, there is a difference between uploading and downloading copyrighted content, which would make it extremely relevant to analyze the extent to which students upload content to peers outside campus or download it from them. Such findings could also be important in terms of quantifying the economic impact of P2P. However, the available data does not allow one to draw significant conclusions regarding downloads versus uploads.

V. RESULTS

ISU is a public college. Initially established as a teacher's college, it still offers many education-related majors. It has about twenty thousand students, of which about 88% are undergraduate and 34%

---

[19] *See* 17 U.S.C. § 107.

[20] Whether or not making copyrighted content available constitutes a copyright violation is currently the subject of legal dispute beyond the scope of this paper (Bangeman, "Judge kills RIAA subpeona: Making available not infringement," http://tinyurl.com/ref-bman-2008b).

live on campus (Illinois State University, 2006). This section presents the results obtained from analysis of the collected data, which focused on students that live on campus and that accounted for over 96% of all detected P2P activity.

A. Extent and Evolution of Detected P2P Activity

Each of the three monitoring periods observed represents part of an academic semester in the one-year period between Spring 2007 and Spring 2008. In Spring 2008, the latter of those monitoring periods, P2P usage, particularly to transfer copyrighted content, remained widespread on campus. As table 2 shows, in Spring 2008 about 40% of students living on campus were observed performing P2P; of these, 70% were detected transferring copyrighted content over P2P, averaging more than four distinct copyrighted movies, songs or TV shows in a day.

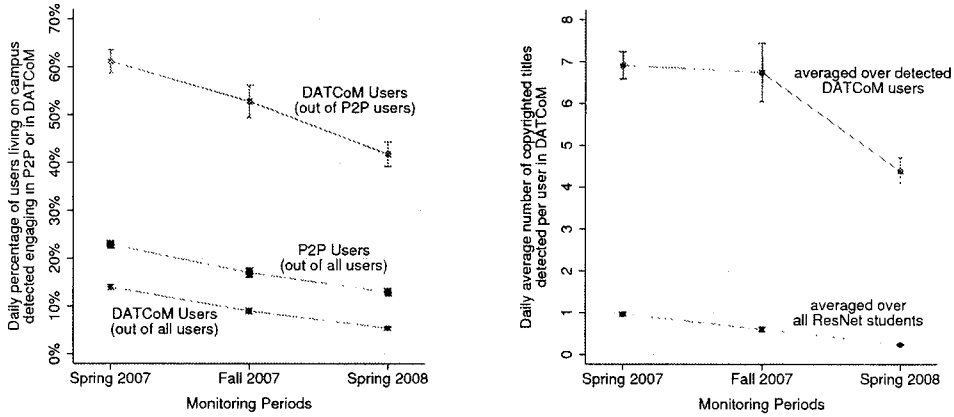
Table 2
Percentage of students detected performing P2P and DATCoMs, and of number of copyrighted titles detected per student in the Spring 2008 monitoring period (95% CI in parenthesis).

	Out of students living on campus	Out of detected P2P users	Out of detected DATCoM users[21]
Students detected in P2P	33.5% (38.3% - 40.7%)		
Students detected in DATCoM	27.6% (26.6% - 28.7%)	70.0% 6(8.2% - 71.7%)	
Copyrighted titles detected per student in the period	7.9 (7.12 - 8.62)	19.9 (18.12 - 21.75)	28.5 (25.99 - 30.96)
Copyrighted titled detected per student per day	0.24 (0.22 - 0.26)	1.82 (1.71 - 1.93)	4.35 (4.16 - 4.55)

Over the one-year period leading to Spring 2008, P2P activity showed a generalized decrease. This is clear from figure 1.a, which shows the declining daily percentage of users detected engaging in

[21] Some titles were detected being shared by some users over several days, therefore the overall number of copyrighted titles detected in the period for each user is not equal to the sum of the number of titles detected in each day.

P2P and transferring copyrighted content, and from figure 1.b, which plots the decrease in the daily average number of copyrighted titles being transferred per student living on campus or per detected DATCoM user. Daily averages allow comparisons of time slices of the same duration for each monitoring period, which makes them adequate for inter-period comparisons.[22] Figures for the entire duration of each monitoring period are not as meaningful for such comparisons due to the different number of monitored hours in each period.



a. b.

Figure 1.

(a) Average daily percentage of students detected engaging in P2P out of all students living on campus, and of students detected engaging in DATCoM, out of all students living on campus and out of detected P2P users in each day.
(b) Daily number of copyrighted media titles detected in DATCoM, averaged over all students living on campus and over students detected engaging in DATCoM in each day. Caps represent 95% confidence intervals (CI).

Hence, despite the observed decrease over time, P2P usage is still widespread on campus, as are transfers of copyrighted content using P2P. The number of P2P users detected on campus in the later Spring 2008 period, while falling below the 50% reported by the RIAA (Lamy et al., 2007), is still in-line with previous survey results reporting widespread use of P2P in other university campuses.

---

[22] In order to draw meaningful comparisons between monitoring periods, and since the period durations are different, we averaged over sub-periods with similar duration of one day, i.e., twenty-four consecutive hours of monitoring data starting at midnight. Some days were excluded from the calculations, namely Spring break and Easter or Labor Day weekend, because the percentage of students present on campus was much lower.

1. INTERPRETING THE RESULTS

While we cannot know the exact reason behind the decrease in the percentage of students detected engaging in P2P and transferring copyrighted content between Spring 2007 and Spring 2008, we can offer various plausible hypotheses to explain this event. One hypothesis is that some students did abandon P2P, either for legal purchases of media or for illegal methods of obtaining content online for free.[23] Another plausible hypothesis is that some students stopped being detected because they activated P2P traffic encryption. Students could abandon P2P or turn to encryption if they knew they were being monitored, and ISU students probably knew that was the case from reading a series of articles about piracy in the ISU student newspaper that warned about plans for network monitoring (Froemling, 2006; Smith, 2006; Swasko et al., 2007). However, given that all such articles were published before Spring 2007, the first monitoring period, we believe this does not explain the observed decline. In fact, during monitoring periods, ISU students had no reason to believe that the likelihood of monitoring increased.

Assuming the possibility that some users turned to encryption, the percentage of P2P users detected on campus should be interpreted as a lower bound because the monitoring appliances used for data collection were unable to classify encrypted P2P traffic as being P2P traffic. Since users that encrypt all P2P traffic will not show up as P2P users in collected data, the precise impact of P2P encryption in the current results depends on whether or not users were knowledgeable of that technique and willing to activate it.[24] Use of technology to limit quality of service or network capacity available for P2P, or to impose some type of punishment on alleged copyright violators, can provide incentives for users to activate evasive measures such as encryption; but the question remains whether such an incentive is enough for users to act. While further technical advances have begun to yield cost-effective network monitoring tools that can detect encrypted P2P

---

[23] Other methods of obtaining content online for free, such as video streaming websites and Direct Download Link (DDL) hosts, were becoming popular at the time of monitoring. While those sites have the potential to divert many users from P2P, from very limited analysis pertaining to use of Rapidshare.com (today's top DDL host) we found that very few users downloaded content from that website. Hence, we believe that DDL hosts were not responsible for the observed decline in P2P usage.

[24] Encryption of P2P traffic is achievable simply by activating a feature available in most modern BitTorrent and Gnutella clients, the two most popular P2P networks currently in use.

traffic, no form of network monitoring can observe the content transferred within such traffic in order to determine whether it is copyrighted or not.

The percentage of DATCoM users and number of copyrighted titles detected per user should also be interpreted as lower bounds because, under several circumstances, appliances used for data collection fail to detect whether content transferred within P2P traffic belongs to copyrighted works.[25]

In conclusion, the data reported above are lower bounds on both general P2P activity and P2P activity related to copyrighted content taking place on campus. The observed decreasing trend may reflect a decrease in actual activity, but may also reflect an increase in use of techniques to conceal online activity from detection via network monitoring.

B. DEMOGRAPHICS OF P2P USERS

Variations of the above figures over demographics (i.e., by gender, class, birth year, area of major[26] or IT savviness[27] of students) and across the three monitoring periods show decreasing trends in the percentage of detected P2P users, percentage of detected DATCoM users, and number of copyrighted titles detected being transferred per user in each demographic subgroup (figures 2 through 5), with only small, but sometimes statistically significant, differences between demographic subgroups. The fact that activity spans all demographics is another way in which P2P and transfers of copyrighted content were and remain widespread on campus.

Differences among subgroups of the campus population could provide insight into possible drivers for P2P use or be useful from an intervention perspective as a means of targeting actions to reduce the volume of transfers of copyrighted content. The relative uniformity of P2P usage and transfers of copyrighted content among demographics not only makes them less useful towards these goals, but also shows that whatever the incentives were for the observed decrease, they

---

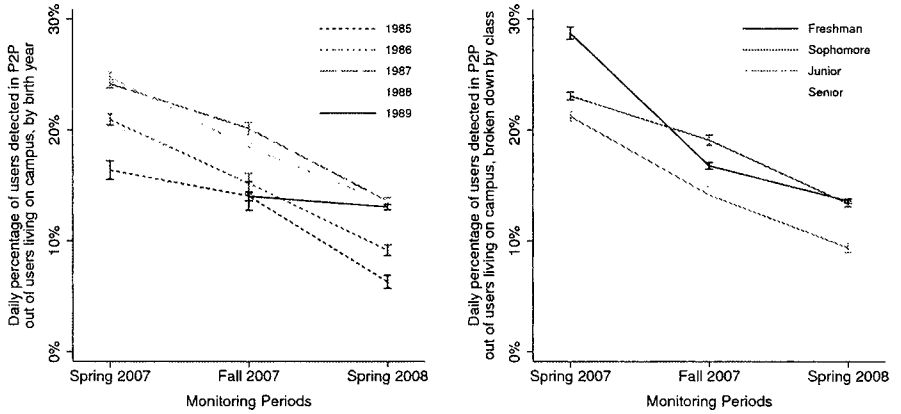
[25] A detailed account of such circumstances is provided by Mateus and Peha "Characterizing digital media exchanges in a university campus network" (2009).

[26] ResNet students were in 79 different majors. Based solely on their major, students were grouped by area of major, which captures the general scientific area of their majors.

[27] IT savviness is inferred from students' majors and it captures the propensity of students to be more IT savvy, which may lead to different online behavior.

seem to have reached all demographics alike. This does not rule out the hypothesis that while responding to the same incentives, students in some demographics turned to measures to conceal their activity while others stopped P2P activity, resulting in the relatively uniform decrease observed.

While not useful for targeting purposes, differences between demographic subgroups can still provide relevant insight about P2P usage on campus. In particular, breakdowns by birth year and by class show that the incidence of detected P2P usage decreases from one monitoring period to the next for every birth year (figure 2.a) and for every class (figure 2.b). Similar decreasing trends are observable in the percentage of users detected engaging in DATCoM as well as in the number of copyrighted titles transferred per DATCoM user. The fact that P2P usage is already high in September for freshman, most of whom have just begun higher education, and that usage declines over time for every given birth year, support claims by higher education officials that students already come to college with entrenched P2P habits and that it is not in college where they "learn" to use P2P. In addition, the fact that freshman in 2008 use P2P less than freshman in 2007, and the same is true for sophomores, juniors, and seniors, indicates that the demand for P2P is fading somewhat with each subsequent "generation" of students.

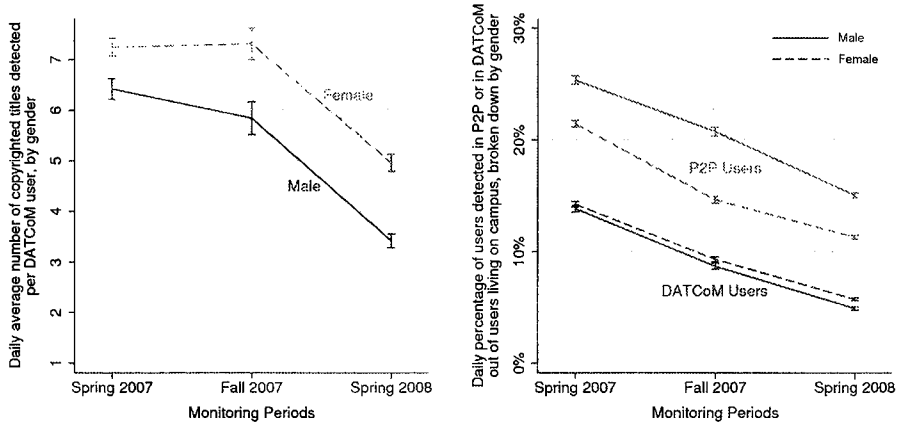


a. b.

Figure 2.

Breakdown of percentage of detected P2P users by birth year and by class. Error bars represent 95% CI. a) Detected P2P users, by birth year: daily percentage of students living on campus detected engaging in P2P broken down by birth year. b) Detected P2P users, by class: daily percentage of students living on campus detected engaging in P2P broken down by class.

Gender and IT Savviness display similar patterns in terms of P2P usage,[28] with a significantly higher percentage of males detected engaging in P2P than females (figure 3.a) and also a significantly higher percentage of IT savvy users than non- IT savvy users (figure 4.a). However, when it comes to the percentage of users engaging in DATCoMs or to the daily average number of copyrighted titles detected being transferred by the latter, then the roles are inverted and females and non- IT savvy users take the lead (figures 3.b and 4.b). This flip can happen because males (or IT savvy students) who engaged in P2P used it to obtain content that either can be legally transferred over P2P or that the monitoring appliances were unable to detect.[29]



a. b.

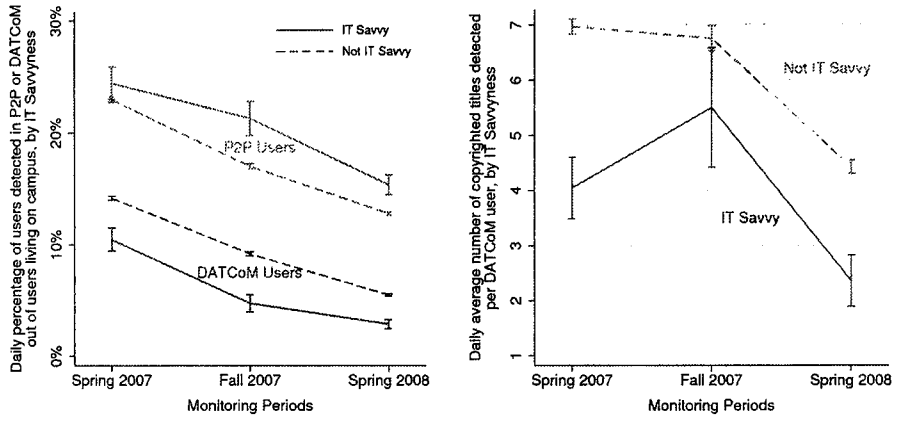
Figure 3.

Breakdown of detected P2P activity by gender. Error bars represent 95% CI. a) Users detected in P2P or DATCoM: daily percentage of students living on campus detected engaging in P2P or in DATCoM. b) Copyrighted titles per DATCoM user: average daily number of copyrighted titles detected being transferred per DATCoM user.

---

[28] The correlation coefficient between gender and IT savviness is 0.17 for all monitoring periods.

[29] Monitoring appliances used can fail to detect specific types of content and users can take measures to prevent the detection of certain types of content. Both cases are further discussed by Mateus and Peha, "Characterizing digital media exchanges in a university campus network," (2009).

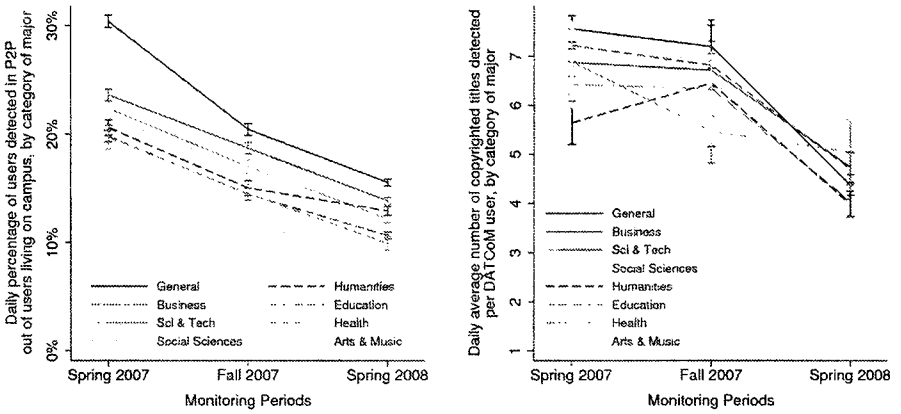


a. b.

Figure 4.

Breakdown of detected P2P activity by IT savviness. Error bars represent 95% CI. a) Users detected in P2P or DATCoM: daily percentage of students living on campus detected engaging in P2P or in DATCoM. b) Copyrighted titles per DATCoM user: average daily number of copyrighted titles detected being transferred per DATCoM user.

Finally, there are few differences in the percentage of students engaging in P2P or DATCoMs among different areas of major. From figure 5.a, out of eight areas of major, detected P2P usage was highest among General Student majors and lowest among Arts and Music majors (similar to users detected in DATCoMs), both with statistically significant difference from other major areas. However, looking at the detected number of titles per DATCoM user (figure 5.b), the differences between majors fade and become insignificant in all periods. Hence, while General Student and Arts and Music majors are respectively more and less prone to engage in detected P2P, those that do transfer copyrighted content tend to transfer as many titles as students in other majors.



a. b.

Figure 5.

Breakdown of detected P2P activity by area of major. Error bars represent 95% CI. a) Detected P2P users, by area of major: daily percentage of students living on campus detected engaging in P2P. b) Copyrighted titles per DATCoM user: average daily number of copyrighted titles detected being transferred per DATCoM user.

C. CONTENT TRANSFERRED OVER P2P

This section presents a characterization of the types of content that students living on campus transferred over P2P, covering both the titles identified as copyrighted and those that could not be identified by the appliances used for monitoring. DPI technology used in detection of copyrighted content can only identify content from a predefined pool of titles. AM in particular can only detect songs, movies, and TV shows because the pool of identifiable content contains features only for titles of these types. This leaves out other types of copyrighted content often found in P2P networks, such as software or adult content. While AM cannot tell whether content of the latter types belongs to titles whose transfer over P2P is not authorized, it can nevertheless observe and report which filenames are exchanged by extracting those filenames from metadata contained within P2P transfers. Analysis in this section uses both media titles contained in DATCoMs as well as filenames contained in communication sessions not identified as DATCoMs.

Overall, in the three monitoring periods AM detected over 36,000 distinct media titles in DATCoMs from about 74% of detected P2P users, and over 100,000 distinct filenames in metadata not identified as copyrighted from about 85% of detected P2P users. As expected, titles detected in DATCoMs were found to be songs, movies, or TV shows, whereas metadata unveils transfers of many other types of

content. Table 3 presents the breakdown of detected titles and filenames by type of content and shows that close to 20% of detected filenames indicate transfers of music albums, adult content, software, books or images.

Table 3

Percentage of copyrighted titles detected in DATCoM and filenames detected in metadata overall in the three periods, and for each individual monitoring period, broken down by type of content. Some titles or filenames were detected in multiple periods. All columns add up to 100%.

	Titles in DATCoM n=36,313	File names in metadata n = 101,879	Titles + Filenames Spring 2007 n = 64,102	Fall 2007 n = 51,770	Spring 2008 n = 46,896
Unclassified		7.7%	3.3%	5.7%	7.8%
Song	99.5%	66.0%	83%	79%	70%
Album		2.3%	0.7%	1.3%	2.8%
Movie	0.4%	3.9%	2.2%	2.4%	3.6%
TV Show	0.2%	3.0%	2.1%	1.5%	2.4%
Other[30]		17.0%	8.7%	11.1%	13.4%

Table 3 also shows a great disparity between the percentage of movies and TV shows detected among DATCoM titles and those detected among metadata filenames. Such disparity is due to success of the content detection mechanisms at identifying copyrighted content within audio files more often than copyrighted content within video files.[31] For this reason, titles detected in DATCoMs and filenames detected using metadata will be considered jointly in the remainder of this section. While filenames may advertise content that is not what is actually stored within the file, P2P searches work by matching keywords to filenames. Therefore, the filenames are still useful for the analysis in this section, which is concerned with figuring out what content P2P users try to get from P2P.

[30] Other stands for: Adult content, Software, Books, Audiobooks and images.

[31] Mateus and Peha, "Characterizing digital media exchanges in a university campus network" (2009).

1. DISTRIBUTION OF POPULARITY OF MUSIC AND VIDEO

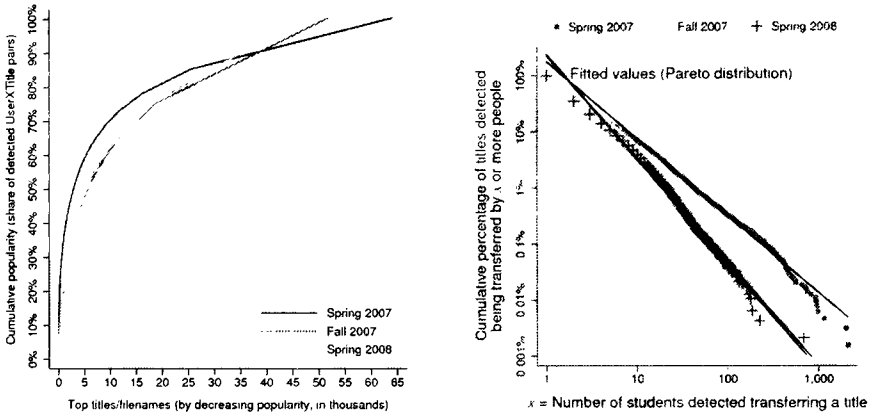
Knowing if P2P is used to transfer the latest blockbuster movie or top-selling single or if it is used to transfer less popular media not available in the store can inform industry decision-makers working on alternate ways to reach P2P users. For example, these alternate ways might include new marketing strategies to make content sales competitive with P2P, or expanding catalogs to make it easy for P2P users to obtain the material they seek from legal sources.

Figure 6.a plots the cumulative distribution of the popularity of titles/files detected being transferred over P2P. The concept of popularity used here is analogous to the one used for actual sales of content: popularity is defined as the percentage of transfers of a title/file (i.e., the percentage of User–Title/Filename pairs involving that title/file) out of the total transfers of all titles/files (i.e., out of the total pool of User–Title/Filename pairs detected) in each period. The distribution shows that a small set of the most popular titles is responsible for most transfers. For instance, in Spring 2007 only 5% of the files/titles represent more than 50% of all transfers. However, it is also clear that P2P caters to less mainstream taste.

The distribution of the number of people detected transferring each title shows a heavy tail of unpopular titles in all monitoring periods. This is clear from figure 6.b, which plots the inverse cumulative distribution of the number of people detected transferring each title/filename in a log-log scale. It shows that the majority of titles/filenames were detected being transferred only by a few students. For instance, in all periods over 90% of detected titles and filenames were transferred by less than 10 students each. The lines in figure 6 result from fitting Pareto distributions ($P[X \geq x] \sim x^{-k}$) to the data in each period.[32] The fitted parameter (k) controls the shape of the distribution: the higher the value of k, the heavier the distribution's tail. In this case, as portrayed in the first column of table 4, k grows from Spring 2007 (1.37) to the later periods (1.84 and 1.87), showing that the tail of the distribution became heavier as time

[32] The shape parameter for the Pareto distribution (k) was estimated using linear regression of the logarithm of the cumulative number of titles transferred by x or more people ($P[X \geq x]$) on the logarithm of the number of people (x), i.e., by fitting linear relationships similar to the ones presented in figure 6.b. The correspondence between the shape parameter of the inverse cumulative Pareto distribution (k) and the exponent of the underlying density power law (a) is given by $a = 1+k$ (Adamic, "Zipf, Power-laws, and Pareto – a ranking tutorial," (2000); Adamic and Huberman, "Zipf's law and the Internet," (2002)).

passed. This indicates that there was an increase in the percentage of titles on the low demand side of the distribution, which hints at a diversification of the media obtained through P2P with P2P users' interest becoming more dispersed through a variety of less popular content and less concentrated in a few very popular titles.



a. b.

Figure 6.

Cumulative distribution function (CDF) of popularity of detected titles/filenames in the different monitoring periods. a) Popularity of top titles: Cumulative share of detected pairs of user x title/filename as a function of the rank of detected titles/filenames (example: in Spring 2007, the top 5,000 Titles/Filenames detected accounted for about 60% of all unique pairs of user x title/filename detected). b) People per title: Markers represent the inverse cumulative distribution of the number of students detected transferring each title/filename, and lines represent the approximation to Pareto distributions (example: in any monitoring period, under 10% of the titles are detected being transferred by ten or more students).

The heavy tail observed in the distribution of popularity of content transferred over P2P shows that P2P can cater to students' interest in a wide variety of media titles. This means that the size of the catalog that is made available to customers is a factor that needs to be taken into account by anyone developing a legal media service if that service is to compete with P2P.

Looking separately at music and video content shows that over the one-year period of monitoring, the distribution of popularity of songs remained relatively stable, while video saw a great shift towards a higher percentage of titles with only a few transfers each. This is visible in table 4, which presents the shape parameters resulting from fitting Pareto distributions separately for music (songs), and video (movies, TV shows, or adult content).

In Spring 2007, the shape parameter for songs (1.76) was greater than for video (0.62 for movies, 0.64 for TV shows, and 1.1 for adult content), meaning that the percentage of files in the low demand region of the distribution was greater for songs than for the other content types, i.e., that students' transfers of videos concentrated more in popular titles than their transfers of music. This is comparable to the distributions of sales of CDs and DVDs from Amazon.com and of movie ratings from Netflix. Estimates for Amazon.com reported sales of CDs and DVDs show shape parameters for a Pareto distribution of 0.99 for CD sales and 0.62 for DVD sales (Smith and Telang, 2007).[33] Using data on title ratings obtained from the Netflix Prize data set[34] yields an estimated shape parameter of 0.66 (R^2=0.88) for Netflix movies.[35] Hence, much like what was observed in campus P2P in Spring 2007, video (whether it is DVD sales or Netflix ratings) seems to be more concentrated around popular titles than music.

Moving forward from Spring 2007 to Spring 2008, we observe a great decline in the number of P2P users detected transferring video[36] (from 50% to 5.5% of P2P users detected transferring movies and TV

---

[33] Actual estimates were based on the distribution of sales per sales rank ($sales \sim rank^{-b}$) and yielded values of b=1.01 for CDs and b=1.61 for DVDs. As demonstrated in Adamic, "Zipf, Power-laws, and Pareto – a ranking tutorial," (2000) and Adamic and Huberman, "Zipf's law and the Internet," (2002), going from estimates of b to estimates of k, the shape parameter of a Pareto distribution ($P[X \geq x] \sim x^{-k}$, which is equivalent to $rank \sim sales^{-k}$), is trivial, since $k = 1/b$.

[34] The Netflix prize dataset (Netflix, "Netflix prize dataset," (2009)) comprises data on over 100 million ratings given by close to 500 thousand users to over 17 thousand movies. The dataset was published at http://www.netflixprize.com as training set for a competition promoted by Netflix and stopped being available for download when the competition ended.

[35] The distribution of popularity of Netflix titles is measured in terms of user ratings and does not reflect actual movie rentals, but it is a fair approximation since movies that are most often rated are also most often rented.

[36] Such a steep decline might be due to some monitoring issue. We considered such a hypothesis, but believe that not to be the case because the monitoring appliances used were the same in all periods. The only thing that changed between periods was the database of copyrighted titles, which is updated periodically with new titles. If it was not properly updated between periods, then more recent (and likely more popular) titles would not be detected as copyrighted. However, that would not influence the reported statistics, which are drawn from both titles identified as copyrighted and filenames detected in transfers that could not be identified as copyrighted, so an outdated database of titles would just shift detection from copyrighted titles to filenames in metadata, maintaining the general overall counts.

shows, and from 27% to 7% transferring adult content). A decline was also observed for music, but it was much smaller (from 66% to 48% of P2P users detected transferring songs). This resulted in fewer people transferring each file/title on average, and in the case of video the decrease was greater for popular titles. As an example, from Spring 2007 to Spring 2008 the number of students detected transferring the most popular movie or TV show at the time of monitoring decreased dramatically from about 1000 to 13 students, and for the most popular adult content file the decrease was from about 650 to 31 students, while the most popular song saw a milder decrease from 430 to 225 students. The result of such decline in terms of the distribution of popularity of transferred titles was that from Spring 2007 to Spring 2008 the distribution of transfers of video changed considerably towards a heavier tail. In Spring 2008, the shape parameter estimate for the distribution of video increased significantly (to 2.53 and 2.54 for movies and TV shows, respectively, and to 2.46 for adult content) while for songs the parameter increased only slightly (to 1.83). As such, the percentage of titles with low demand increased for both types of video but remained roughly the same for songs. This means that transfers of video over P2P started to dissipate over a wider range of titles instead of concentrating in a few popular titles, which once again highlights the importance of a large video catalog in a legal service that aspires to compete with P2P.

Table 4.

Estimates of Pareto distribution shape parameters obtained by fitting the distribution of popularity of song, movie, TV show, and adult titles and filenames detected being transferred in each monitoring period. Each cell contains point estimates, 95% confidence intervals in parenthesis, and adjusted R^2 values.

Percentage of students detected performing P2P and DATCoMs, and of number of copyrighted titles detected per student in the Spring 2008 monitoring period (95% CI in parenthesis).

	All Media	Songs	Movies	TV Shows	Adult
Spring 2007	-1.37 (-1.39 to -1.35) R^2=0.988	-1.76 (-1.81 to -1.72) R^2=0.978	-0.62 (-0.89 to -0.55) R^2= 0.805	-0.64 (-0.69 to -0.60) R^2=0.922	-1.10 (-1.15 to -1.06) R^2=0.955
Fall 2007	-1.84 (-1.88 to -1.79) R^2=0.982	-1.90 (-1.95 to -1.86) R^2=0.985	-3.06 (-3.30 to -2.82) R^2=0.993	-3.00 (-3.19 to -2.81) R^2=0.995	-2.53 (-2.73 to -2.33) R^2=0.970
Spring 2008	-1.87 (-1.91 to -1.83) R^2=0.988	-1.83 (-1.88 to -1.79) R^2=0.988	-2.53 (-2.84 to -2.23) R^2=0.969	-2.54 (-2.82 to -2.27) R^2=0.978	-2.46 (-2.69 to -2.22) R^2=0.948

Many factors might have caused the observed change in the video distribution, in particular when compared to the relative stability of the song distribution, none of which can be readily observed with the available data. One possible factor is the emergence of other sources of media, namely one-click hosts and video streaming websites, which increased in popularity during the monitoring periods. However, while we have insufficient data to exclude this possibility entirely, these online sources do not appear to be the cause. For one-click hosts and streaming websites to account for the shape change in the distribution of popularity for video, P2P users would need to be obtaining popular videos from those alternatives and using P2P only for less-popular videos. Although this is possible, there are arguments against this hypothesis. One-click hosts provide music as well as video. Given the stability in the music distribution, users would need to have downloaded popular video from the one-click host but not music. Streaming websites are devoted to video, so the hypothesis only makes sense if users downloaded popular titles from those websites and downloaded unpopular video from P2P. This is plausible, but we do not have an account of what content was available

or downloaded from one-click hosts or streaming websites at the time, so we cannot reach a conclusion in this case.[37]

Another possibility concerns the maturity of the P2P "market" for music, which has existed since 1999 and was not subject to much change over the course of the year, versus the still embryonic P2P "market" for video, which continues to evolve rapidly. In particular, there were two significant trends that might have affected how people obtain video from P2P. On one side, the amount and variety of video content available in P2P (in particular through BitTorrent, the main P2P network used for video transfers) increased significantly during the monitoring period, which allowed users to more easily find the content they desired and gave them a wider choice of video content to transfer, thus catering to more diverse user preferences. On the other side, the bandwidth available to Internet users was continually expanding, making it less time-consuming to get large video files from P2P. These trends resulted in a decrease in the "cost" of obtaining a video title from P2P, where cost is measured in terms of the search time necessary to find the desired title in a P2P network (which decreases as more titles become available), plus the actual transfer time of the video file if the title is available (which decreases if there is more bandwidth available for the transfer). So, the fact that it was more expensive to obtain a video from P2P in Spring 2007 than in the subsequent periods could result in users sticking to only a few titles they really wanted in the earlier period, which yields a distribution more concentrated around popular content; while in later periods users could more freely experiment with a wider variety of titles, which leads to a heavier tailed distribution of transferred content.

2. OTHER CONTENT

P2P file sharing is used for multiple purposes, some of which are legal. Depending on the monitoring period, between 20% and 30% of users detected using P2P did not generate DATCoMs. It is unknown how many of these students transferred copyrighted material without being detected and how many were involved only in legal use of P2P. This section tries to understand whether legal uses of P2P, or uses that may not be legal but that would not generate DATCoMs, can explain why so many P2P users did not generate DATCoMs. To do so, it

---

[37] Certainly, this hypothesis would be better grounded if Hulu.com existed at the time, since it typically offers only a few recent TV episodes, but Hulu.com only opened to the general public in March 2008, six months after the observed change in the popularity distribution for video.

focuses on particular content that would not generate DATCoMs when transferred using P2P, namely the Linux OS, the game World of Warcraft (WoW), and adult content, which were identified using metadata.

Transfers of Linux and other open source software are among the most often mentioned legal uses of P2P and generate no DATCoMs. In all of the monitoring periods, the percentage of P2P users transferring files whose names indicate Linux[38] topped at about 0.2%. The breakdown of this percentage between users with detected DATCoMs and users without detected DATCoMs reveals proportions that are not significantly different from zero. Hence, there is no evidence to support the hypothesis that the transfer of Linux is a motivator for the use of P2P, even among users not detected using P2P to transfer copyrighted material.

World of Warcraft (WoW) is the most popular online multi-player game with a market share of over 60% and over 11 million active users[39] and enjoys great popularity among college students. The game producers released patches to improve the game's functionality in all monitoring periods.[40] Game patches are distributed using the BitTorrent protocol,[41] and thus WoW players are likely to be detected as P2P users because they transfer WoW patches after each update. Transfers of WoW patches were detected in all monitoring periods (decreasing from about 7% of detected P2P users in Spring 2007 to 4% in Spring 2008, and averaging at 4.4% of detected P2P users over all periods). The percentage of students detected transferring WoW but not detected in DATCoMs increased over time (from under 1% of P2P users in Spring 2007 to about 3% by Spring 2008) but remained lower than that of students detected transferring WoW alongside copyrighted music and video when accounting for the three monitoring periods altogether (1.4% of P2P users detected

---

[38] Files with names that include "linux" or specific major Linux distributions on the market: "fedora," "ubuntu," "suse," "red hat," "mandriva," "slackware," or "debian."

[39] Magrino, "World of Warcraft hits 11 million," (2008).

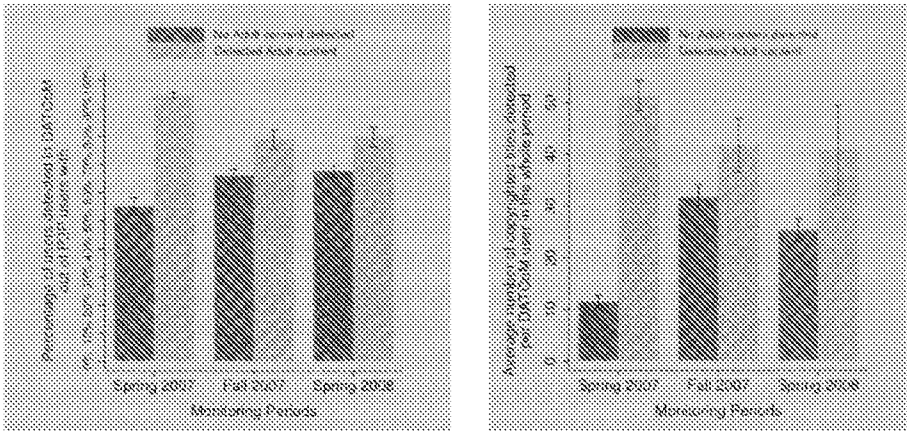
[40] WoW patch releases in the U.S.: v2.0.12 (April 3, 2007), v2.2.0 (September 25, 2007), v2.4.0 (March 25, 2008), and v2.4.1 (April 1, 2008).

[41] The game updater application, which is responsible for implementing the patch download mechanism, relies on the BitTorrent protocol to download game patches. This application is not the same as the one students would use to obtain copyrighted movies and music, although both applications generate traffic that is classified by DPI under the BitTorrent class.

transferring WoW alone, versus 3% detected transferring WoW and engaging in DATCoMs). These figures show that while there are students who play WoW and are detected doing P2P because they download patches for WoW, the greatest part of students that transfer WoW patches also use P2P to obtain copyrighted music and video.

All monitoring periods show significant percentages of P2P users transferring files whose filenames indicate adult content, although with a great decrease from a daily average of 27% of detected P2P users in Spring 2007 to 7.3% in Spring 2008.[42] Transfers of adult content did not result in DATCoMs because the database of copyrighted titles used for identification did not include features for this type of content in any of the monitoring periods. This means that users who transfer adult content but no copyrighted music or video can be detected as P2P users but not as DATCoM users. Nevertheless, detected transfers of adult content do not explain the percentage of users found engaging in P2P but not in DATCoMs since between 80% and 90% of users detected transferring adult content are still detected in DATCoMs (figure 7.a). In fact, DATCoM users detected transferring adult content are observed transferring more copyrighted titles on average than DATCoM users not detected transferring adult content, as shown in figure 7.b.

[42] One possible explanation for the abrupt decrease in the percentage of P2P users detected transferring adult content is the growing availability of websites that offer free streaming of adult content, some of which is user-generated (Lyn, "The deal with user-generated content," (2008)). Such websites are pointed out by the Adult movie industry as being responsible for a decrease in sales of adult movies, an effect that is also expected in P2P transfers of Adult movies.



a. b.

Figure 7.

Breakdown of the percentage of DATCoM users and of the number of transferred titles, by whether users were detected transferring adult content or not. a) Percentage of DATCoM users: The percentage of P2P users detected in DATCoM overall in each monitoring period. b) Number of titles per DATCoM user: The average number of copyrighted titles detected per DATCoM user overall in each monitoring period.

D. RELATIONSHIP BETWEEN USAGE OF P2P AND ITUNES

The impact of unauthorized P2P transfers of copyrighted content on the revenues of copyright holders is partly dependent on how much these transfers displace sales of content that could otherwise happen. While our data does not allow assessment of the sales of copyrighted content that fail to happen due to P2P, it allows comparisons of media-related activity from popular online media outlets in order to assess whether the behavior towards obtaining media from legal online sources differs between users who perform P2P and users who do not perform P2P. This section compares media-related activity from P2P and the iTunes Store (iTS) in Spring 2008 to assess whether P2P users also obtain media from the iTS and whether there are differences between download of free content (song samples) and content that is paid for (songs or videos).

1. METHODOLOGY

This analysis compares P2P and iTunes activity for each IP address in the ISU network. It uses the communication sessions that Packeteer detected for each IP address that were classified as iTunes traffic or as P2P traffic. In this case we believe that the activity detected for IP addresses is very likely to have been performed by a

single network user; the reason for that is that IP addresses in ISU's residence halls were leased statically for the entire semester to each user who acquired a dorm room Internet connection. This allows accounting separately for each user connected to the network in dorm rooms.

However, collected data is composed of activity detected for all IP addresses on campus, some of which are not guaranteed to map to a single user during the entire monitoring period. For that reason, pre-processing had to be conducted in order to exclude two groups of detected IP addresses from analysis: addresses for which the Network Address Translation (NAT) could not be reversed and addresses corresponding to short DHCP leases. The first group exists because Packeteer was deployed on the Internet side of the NAT at the border of the campus network, and thus all IP addresses from within campus were detected in the form of ISU-external translated addresses. This required post-processing at the time of data collection to reverse such translation using the NAT logs, which was not always successful. As a result, 28% of all detected communication sessions were disregarded in the analysis because they could not be attributed to ISU-internal IP addresses, which made the sessions impossible to attribute to single particular devices.[43] The second group corresponds to IP addresses with short DHCP leases. These were not considered for analysis because they do not capture the behavior of single network users, but most likely of several users assigned to each particular address over time.[44] Such addresses amounted to about 9% of all detected addresses, but they represent negligible network activity relevant to the analysis in this section (8MB of traffic per address on average in the entire period and percentages detected with either P2P or iTunes traffic not statistically different from zero).

---

[43] NAT occurs at the IP level in the protocol stack, which makes it independent from anything higher in the stack, particularly transport protocols or application protocols. The ability to translate back using NAT logs maintains this independence. Hence, translation failures are not related to the type of activity contained in the events, which makes us expect these 28% bytes to be missing uniformly across protocols, thus not biasing results towards any type of activity in particular.

[44] Short DHCP leases were not very common in the ISU campus, occurring mostly on the wireless network, which was only available in few places on campus. DHCP-leased IP addresses appear in the data either only once for a short period of time (when the IP address is leased a single time in the monitoring period) or several times, but never consecutively for more than the duration of the DHCP lease period (for IP addresses which are recycled by the DHCP server, and therefore leased multiple times in the monitoring period).

In the case of iTunes communication sessions, to separate transfers of different types of media from the iTS, each detected inbound iTS communication session was classified into a media category using the following criterion:[45] sessions with less than 480KB were considered control traffic, sessions with 480KB to 1MB were considered sampling of music, sessions with 1MB to 25MB were considered downloads of songs, and sessions with more than 25MB were considered downloads of video.[46] Given that an iTunes user can generate traffic without entering the iTS[47] and analysis concerns song and video-related activities in the iTS, IP addresses were considered to have used the iTS if at least one sampling, song, or video activity was detected. Such addresses correspond to 41% of addresses with detected iTunes traffic.

In the case of P2P communication sessions, since the focus of analysis is in the activity of P2P users that likely transferred some copyrighted media from a P2P network, only IP addresses detected transferring enough P2P bytes to constitute a copyrighted title are considered (a threshold that is set at 3 MB, roughly the amount of traffic necessary to transfer one song).

2. RESULTS

Using the above criteria, IP addresses are classified as P2P or iTS users. Simple tabulations on the data show that use of P2P is found to be correlated with use of the iTS. As shown in table 5,[48] IP addresses

---

[45] This criterion was defined based on observation of the distribution of bytes per inbound communication session with detected iTunes traffic. That distribution displayed clear peaks around traffic volumes that indicate specific activities: around 480KB and around 960KB, equivalent to 30 seconds of a song at a bitrates of 128kb/s and 256kb/s respectively, likely corresponding to music sampling activities; centered around 4MB, likely corresponding to downloads of songs; and above 25MB, with a clear peak around 500MB, likely corresponding to downloads of videos.

[46] There are clearly other types of media that can be acquired from the iTunes Store, such as podcasts or iPod games (in Spring 2008 iPhone App store did not exist yet, hence there are no iPhone application transfers in the monitored events). We assume that the percentage of students that access these types of content was small.

[47] An example of such traffic is the download of album covers when the user transfers music from a CD to her iTunes music library.

[48] The percentages in table 5 refer to IP addresses with or without detected P2P activity. They are different from the percentage of students detected engaging in P2P reported in section V.A because they refer to different entities, as made clear in the methodology subsection above.

detected engaging in P2P are more likely to be detected using the iTS than IP addresses not detected engaging in P2P (and vice-versa). Clearly, to some extent, P2P and the iTS complement each other, a fact that can impact the revenues of copyright holders whose content is sold in the iTS in different ways. If a student uses P2P only when the content she is seeking is not available on the iTS, then P2P transfers of that content have no impact on iTS sales. On the opposite side, if a student uses the iTS only to sample content that she then gets from P2P, then all revenue from eventual iTS sales is lost.[49] Between the two extremes fall students who buy some content from the iTS and who get some content for free from P2P.

To investigate the relationship between content sampling and purchasing from the iTS and P2P usage, table 5 also breaks down iTS users between those who only sampled content and those who actually purchased content.[50] The table shows that about one third of P2P users still use the iTS, and that close to one quarter of P2P users still purchase content from the iTS, which means that while use of P2P may reduce the number of people who purchase from the iTS, it certainly does not eliminate it. Thus, not all iTS revenue is lost to P2P. Furthermore, P2P users are about three times more likely to use the iTS than non-P2P users, but they are four times more likely to use it only for sampling than non-P2P users. This means that P2P users are somewhat more likely to use iTS resources only to decide what content to obtain, possibly from P2P.

---

[49] The fact that a user samples content from the iTS and then transfers all that content from P2P does not necessarily mean that all the titles transferred from P2P are lost sales. Due to budget constraints or due to willingness to pay for some content being below the price of that content, it is possible that the user would not acquire all the sampled content if she had no way to get it for free.

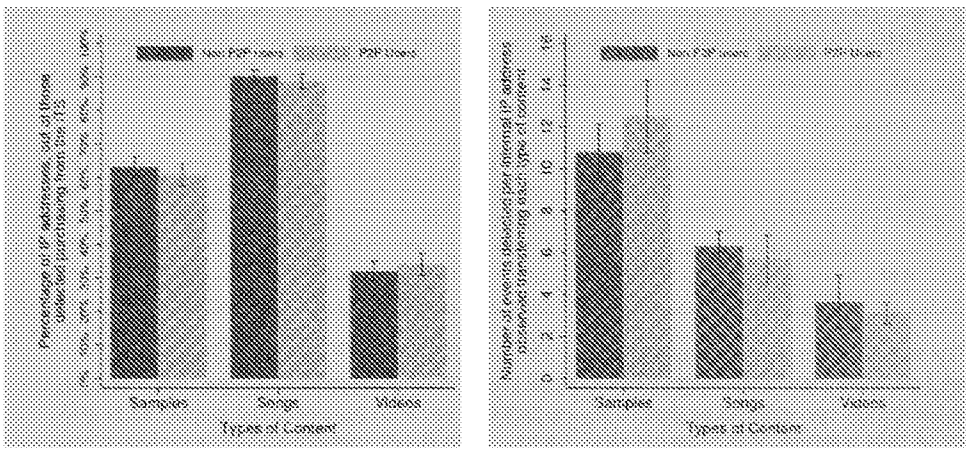
[50] Table 5 presents the percentages of IP addresses engaged in each activity. Although this is not identical to the percentage of users engaged in this activity, it is similar enough for us to reach conclusions. We believe the primary difference will come from the IP addresses of devices not permanently operated by a user, such as servers, which do not engage in P2P or iTS activity. However, in order for those addresses to have an influence that would alter our general conclusions they would need to comprise over 40% of all observed IP addresses, which does not seem plausible on a campus with twenty thousand students.

Table 5.

Cross tabulations of detected P2P with detected iTS activity (broken down by iTS users detected only transferring content samples or detected transferring songs or videos). Percentages in columns add up to 100%

	Not P2P (76.4%)	P2P (23.6%)
Did not use the iTS (84%)	89.1%	67.4%
Used the iTS only to sample content (4.6%)	2.7%	10.6%
Used the iTS to purchase songs or videos (11.4%	8.2%	22.0%

Focusing on users who purchase content from the iTS at some point in the monitoring period, there is no statistically significant difference between P2P and non-P2P users when it comes to the percentage of users purchasing songs or videos, or the number of songs or videos purchased per user. As figure 8.a shows, about 90% of the users who purchase content from the iTS purchase songs and over 30% purchase videos, equally among P2P users and non-P2P users. Furthermore each P2P user who buys songs (or videos) buys as many songs (or videos) on average as each non-P2P user, as depicted in figure 8.b.



a. b.

Figure 8.

Percentage of IP addresses and average number of events per IP address detected in the iTS, broken down by type of content. (a) Percentage of IP addresses detected sampling music, downloading songs and downloading videos out of those detected using the iTS to purchase content, broken down by P2P usage. (b) Average number of samples, songs and videos downloaded per IP address detected downloading each of such media from iTS, broken down by P2P usage. Caps represent 95% confidence intervals.

It is not possible to tell precisely what impact P2P has on paid services from the detected activity alone, but there are certainly things to learn. There is evidence consistent with the hypothesis that some P2P users use iTS only to obtain free samples, since, out of iTS users, those who do P2P are somewhat more likely to use the iTS only to sample content for free without purchasing than those who do not do P2P. On the other hand, there is evidence that contradicts the hypothesis that P2P users view the ability to transfer content for free as a complete substitute for paid services, since a substantial fraction of P2P users also purchase content from the iTS. Moreover, purchasing behavior in the iTS is very similar for P2P and non-P2P users, i.e., P2P users who purchase content from the iTS do it in comparable percentages and download comparable quantities of songs and videos as non-P2P users who purchase content from the iTS.

VI. CONCLUSIONS AND POLICY IMPLICATIONS

Using data collected between Spring 2007 and Spring 2008, we found that P2P activity and use of P2P to transfer copyrighted content were widespread on a university campus. In Spring 2008, the latest monitoring period, 40% of the students living on campus were observed using a P2P protocol. Moreover, while P2P can certainly be used effectively for legal purposes, there is evidence that many P2P users are violating copyright law. Seventy percent of those observed using P2P were observed attempting to transfer copyrighted material, and most such attempts are likely to be copyright violations. On average users detected transferring copyrighted material were observed attempting to transfer four distinct copyrighted media titles during a day of monitoring. Since not all such activity is observable with the Deep Packet Inspection (DPI) tools employed, the actual number of copyrighted titles transferred and the actual percentage of students engaged in these transfers is likely to be higher. Although some people might use P2P only to transfer content legally, we found no evidence that this was common. We considered transfers of the Linux operating system, of World of Warcraft (WoW) game patches, and of adult material (whose transfer over P2P may be legal for some titles but not for others), and found no evidence to support the hypothesis that a significant fraction of students use P2P to obtain those types of content and not to obtain copyrighted music and video as well.

The fact that laws are violated so frequently is good reason to consider a significant change. This could take the form of a change in policies, business practices, enforcement methods, technology, consumer education, or a combination of these. Although our results alone cannot tell us exactly what approach should be followed, they can help answer some important questions.

One such question is whether any particular intervention, such as a campaign to educate Internet users about copyright law, should be targeted at a specific group or applied generally to all. At least among college students, we found no reason to target any specific group because P2P users and users transferring copyrighted content were detected across all demographics, with fairly similar incidence among different genders, ages, classes, and majors. However, if education campaigns prove to be useful, there is reason to reach students even before they get to college. We find high incidence of P2P activity for freshmen in their first month on campus, followed by a gradual decrease in subsequent semesters. This indicates that students became P2P users prior to entering college, which is consistent with claims by higher education officials that students "learn" to use P2P at a younger age and already come to college with entrenched P2P habits.

Another question concerns whether and how the industries that produce and distribute copyrighted content could respond to the challenge of competing with a free but illegal alternative. Although it is impossible to quantify from our data how the number of transfers of copyrighted titles observed translates to lost sales, as this depends on many factors, it is reasonable to assume that some sales are lost. Nevertheless, contrary to the common belief that P2P users simply replace paid services with unpaid, our comparison of P2P usage to usage of the iTunes Store (iTS) showed that 22% of detected P2P users still made purchases from the iTS. Moreover, P2P users were more interested in content from the iTS than non-P2P users and did a significant amount of purchasing; each P2P user who purchased from the iTS purchased on average as much as each non-P2P user who purchased from the iTS. The fact that close to one quarter of P2P users decide to pay for content at times and to get content for free from P2P (risking legal action by copyright holders) at other times means that there must be factors other than price or risk of legal action weighing in such users' decision.

The variety of content available in each outlet appears to be one such factor. In order to offer its customers a range of content that is comparable to what we observed with P2P, a legal media distribution service would have to offer its users a large selection. We found that a small share of high demand titles accounted for most detected P2P

transfers on campus, but also that a heavy tail of low demand titles added up to a significant share of transfers. This means that users' preferences go beyond the most popular titles, and that P2P gives such users access to an extensive catalog to suit those preferences. Unless legal alternatives also provide a large collection of content from which users can choose, users may turn to illegal forms of obtaining the content they desire. While providing such a large selection of content may be possible for online services, it is typically not possible for brick and mortar providers of paid content (CDs, DVDs) who are left at an inherent competitive disadvantage. It is probably worth exploring other factors that might influence a user's decision to choose P2P or a legal paid service for copyrighted content. Factors to consider include the convenience or ease of use of each service, ease of search, transfer speed, and usage restrictions in obtained content. In this way, those legally distributing content online may perhaps make their offerings more attractive in order to compete with the free but illegal alternative, at least for the roughly one quarter of P2P users who are still purchasing some media. This may reduce the impact of P2P use on the revenues of copyright holders; although considering that more than three quarters of P2P users were not seen purchasing from the iTS, at the time the top music retailer in the U.S.,[51] it probably would not eliminate that impact.

Finally, by looking at the evolution of P2P activity over time, we find what might be the beginning of a shift away from observable P2P. Although our observation of only three points in time does not allow for a strong conclusion, we did observe P2P activity on campus decreasing, both in terms of users detected using P2P (10% decrease) and of users transferring copyrighted material (20% decrease, out of detected P2P users), as well as in the average number of unique copyrighted titles detected being transferred per user in a day of monitoring (decreasing from seven to four titles per user detected transferring copyrighted content). Despite this decrease, the detected percentage of P2P users is still consistent with numbers reported by the entertainment industry.[52]

Such a decline can mean that students abandoned P2P for legal services or for other methods of obtaining content online that can also violate copyright law, such as video streaming websites, direct download link providers (e.g., RapidShare), or Usenet newsgroups.

---

[51] Neumayr and Roth, "iTunes Store top music retailer in the U.S.," (2008).

[52] Lamy, Duckworth, and Kennedy, "RIAA launches new initiatives targeting campus music theft," (2007).

Copyright holders clearly benefit from the continuing growth in use of legal options, but emerging illegal sources also matter, and the extent of their impact is an important open question that may affect the future of online copyright enforcement.

The observed decline can also be due to students escaping detection by using methods to conceal their P2P activity from network monitoring. This would have important implications for anyone considering use of technology such as DPI to deter copyright violations. As with any technology, those considering deployment must determine whether intended benefits outweigh costs and any unintended side effects. Whether user behavior would ultimately reduce a system's ability to detect transfers of copyrighted material is certainly one of the factors to consider. DPI has limitations that can preclude it from detecting P2P activity and activity involving copyrighted content under certain conditions,[53] one of which is if P2P users activate encryption, which is available in today's most popular P2P clients. It remains a question whether or not users will exploit these limitations to evade detection. Our future work will investigate this use of encryption, and more generally describe a number of possible technological approaches to detecting online copyright violations, and their respective advantages, disadvantages, and policy implications.

VII. ACKNOWLEDGEMENTS

The authors thank the financial support provided by Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal Program under Grant SFRH/BD/27350/2006 and by the Information and Communication Technologies Institute (ICTI) project entitled "Web Security and Privacy: Weaving Together Technology Innovation with Human and Policy Consideration" (CMU-PT/SE/0028/2008). The authors also thank Illinois State University and in particular the Digital Citizen Project for their cooperation in facilitating the collection of data.

---

53 Limitations in DPI monitoring are further detailed in Mateus and Peha, "Characterizing digital media exchanges in a university campus network," (2009).

I. REFERENCES

105th Congress. 1998. *Digital Millennium Copyright Act (DMCA). Pub. L. 105-304.* October 28.

Adamic, Lada A. *Zipf, Power-laws, and Pareto - a ranking tutorial.* Online tutorial. HP Labs, Information Dynamics Lab. http://www.hpl.hp.com/research/idl/papers/ranking/.

Adamic, Lada A., and Bernardo A. Huberman. 2002. Zipf's law and the Internet. *Glottometrics* 3: 143-150. http://www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf.

Andersen, Birgitte, and Marion Frenz. 2008. *The Impact of Music Downloads and P2P File-Sharing on the Purchase of Music: A Study for Industry Canada.* University of London Working Paper.

Anderson, Nate. 2008. UK ISPs don't want to play umpire to "three strikes" rule. *Ars Technica.* February 15. http://tinyurl.com/ref-anderson-2008.

———. 2009a. French court savages "three-strikes" law, tosses it out. *Ars Technica.* June 10. http://tinyurl.com/ref-anderson-2009.

———. 2009b. France govt goes into spin mode to salvage three-strikes law. *Ars Technica.* June 11. http://tinyurl.com/ref-anderson-2009b.

Bangeman, Eric. 2007. New bill would punish colleges, students who don't become copyright cops. *Ars Technica.* November 11. http://tinyurl.com/ref-bman-2007.

———. 2008a. France's plan to turn ISPs into copyright cops on track. *Ars Technica.* January 28. http://tinyurl.com/ref-bman-2008c.

———. 2008b. Judge kills RIAA subpoena: making available not infringement. *Ars Technica.* April 3. http://tinyurl.com/ref-bman-2008b.

Bhattacharjee, Sudip, Ram D. Gopal, Kaveepan Lertwachara, James R. Marsden, and Rahul Telang. 2007. The Effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts. *Management Science* 53, no. 9: 1359-1374.

Bremner, Charles. 2008. France to ban illegal downloaders from using the Internet under three-strikes rule. *The Times.* June 19.

http://tinyurl.com/ref-bremner-2008.

Buskirk, Eliot Van. 2007. A Poison Pen From the RIAA. *Wired.* February 28. http://tinyurl.com/ref-buskirk-2007.

EFF. 2008. *RIAA v. The People: Five years later.* Electronic Frontier Foundation.

Fischer, Ken. 2007. Bill would force "top 25 piracy schools" to adopt anti-P2P technology. *Ars Technica.* July 23. http://tinyurl.com/ref-fischer2007.

Froemling, Todd. 2006. ISU develops new file sharing options. *Daily Vidette at Illinois State University*, May 2. http://is.gd/eFciC.

Gopal, Ram D., and Sudip Bhattacharjee. 2006. Do Artists Benefit from Online Music Sharing? *Journal of Business* 79, no. 3: 1503-1533.

Guess, Andy. 2008. Downloading by Students Overstated. *Inside Higher Ed.* January 3. http://tinyurl.com/ref-guess-2008.

Hong, Seung-Hyun. 2007. Measuring the Effect of Digital Technology on the Sales of Copyrighted Goods: Evidence from Napster. *Available at SSRN.*

Illinois State University. 2006. *ISU FactBook 2006-2007.*

———. 2008. Digital Citizen Project at Illinois State University, Summary of Project. *Digital Citizen Project at Illinois State.* February. http://www.digitalcitizen.ilstu.edu/summary/.

Lamy, Jonathan, Cara Duckworth, and Liz Kennedy. 2007. *RIAA Launches New Initiatives Targeting Campus Music Theft.* News Release. The Record Industry Association of America, February 28. http://tinyurl.com/ref-riaa-2007.

———. 2008. *RIAA Continues College Deterrence Campaign Into 2008.* News Release. The Record Industry Association of America, January 10. http://tinyurl.com/ref-riaa-2008.

Liebowitz, Stan J. 2008. Research Note--Testing File Sharing's Impact on Music Album Sales in Cities. *Management Science* 54, no. 4: 852-859. doi:10.1287/mnsc.1070.0833.

Lyn, Jett. 2008. The Deal with User-Generated Content. *Xbiz.com.* April 12. http://www.xbiz.com/articles/92416.

Macavinta, Courtney. 1999. Recording industry sues music start-up, cites black market. *Cnet News*. December 7. http://tinyurl.com/ref-mac-1999.

Magrino, Tom. 2008. World of Warcraft hits 11 million. *Gamespot.com*. October 28. http://tinyurl.com/ref-magrino-2008.

Mateus, Alexandre M., and Jon M. Peha. 2009. Characterizing digital media exchanges in a university campus network. 37th Research Conference on Communication, Information and Internet Policy (TPRC). http://www.ece.cmu.edu/~peha/campus_media_exchanges.pdf.

McBride, Sarah, and Ethan Smith. 2008. Music Industry to Abandon Mass Suits. *The Wall Street Journal*, December 19. http://online.wsj.com/article/SB122966038836021137.html.

Michel, Norbert J. 2006. The Impact of Digital File Sharing on the Music Industry: An Empirical Analysis. *Topics in Economic Analysis & Policy* 6, no. 1.

Netflix. 2009. Netflix prize dataset. *Netflixprize.com*. http://www.netflixprize.com/.

Neumayr, Tom, and Jason Roth. 2008. iTunes Store Top Music Retailer in the US. April 3. http://www.apple.com/pr/library/2008/04/03itunes.html.

Oberholzer-Gee, Felix, and Koleman Strumpf. 2007. The Effect of File Sharing on Record Sales: An Empirical Analysis. *Journal of Political Economy* 115, no. 1: 1-42. doi:doi:10.1086/511995.

———. 2009. *File-Sharing and Copyright*. Working Paper 09-132. Harvard Business School.

Oster, Seth. 2008. *MPAA Statement on Motion Picture Industry Losses due to Piracy among College Students*. News Release. MPAA, The Motion Picture Association of America, January 22.

Oswald, Ed. 2006. RIAA Sues LimeWire Over Piracy. *Betanews.com*. August 4. http://tinyurl.com/ref-oswald-2006.

Packeteer. 2007. Applications, Protocols, and Services Classified by PacketWise 7.3. https://bto.bluecoat.com/packetguide/7.3/reference/services.htm.

Pfanner, Eric. 2009. France Approves Crackdown on Internet Piracy.

The New York Times, May 12. http://www.nytimes.com/2009/05/13/technology/internet/13net.html.

RIAA. 2007. *Piracy Online*. News Release. The Recording Industry Association of America. http://tinyurl.com/ref-riaa-2007b.

Rob, Rafael, and Joel Waldfogel. 2006. Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students. *The Journal of Law and Economics* 49, no. 1: 29-62. doi:doi:10.1086/430809.

Smith, Jamie. 2006. ISU staff speaks in Washington on issues of campus piracy. *Daily Vidette at Illinois State University*, October 3. http://is.gd/eFbpI.

Smith, Lamar, George Miller, Howard McKeon, Howard Berman, and Howard Coble. 2007. *Letter sent to universities on May 1, 2007*. http://tinyurl.com/ref-letter-2007.

Smith, Michael D., and Rahul Telang. 2007. *Research Note: Internet Exchanges for Used Digital Goods: Empirical Analysis and Managerial Implications*. Research Note. Tepper School of Business, Carnegie Mellon University.

———. 2008. *Competing with free: The impact of movie broadcasts in DVD sales and Internet piracy*. Working Paper. Carnegie Mellon University. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1028306.

Swasko, Mick, Deanne St. John, Eric Strand, and Mary Yurgil. 2007. Being a pirate is against the law. *Daily Vidette at Illinois State University*, February 21. http://is.gd/eFbUA.

Tanaka, Tatsou. 2004. Does file sharing reduce CD sales?: A case of Japan. In *Conference in IT Innovation*.

U.S. Congress. House. 2003. *Peer-to-Peer Piracy (P2P) on University Campuses*. Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 108th Cong. 1st sess. SN 2, February 26. http://purl.access.gpo.gov/GPO/LPS59277.

———. 2004. *Peer-to-Peer Piracy (P2P) on University Campuses: An Update*. Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 108th Cong. 2nd sess. SN 112, October 5. http://purl.access.gpo.gov/GPO/LPS58080.

———. 2005. *Reducing Peer-to-Peer (P2P) Piracy on University Campuses: A Progress Update.* Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 109th Cong. 2nd sess. SN 109-56, September 22. http://purl.access.gpo.gov/GPO/LPS66466.

———. 2006. *The Internet and the College Campus: How the Entertainment Industry and Higher Education are Working to Combat Illegal Piracy.* Committee on Education and the Workforce. Subcommittee on 21st Century Competitiveness. 109th Cong., 2nd sess. SN 109-58, September 26. http://purl.access.gpo.gov/GPO/LPS76269.

———. 2007a. *An Update: Piracy on University Networks.* Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 110th Cong., 1st sess., SN 110-29, March 8. http://purl.access.gpo.gov/GPO/LPS85994.

———. 2007b. *The Role of Technology in Reducing Illegal File-sharing: A University Perspective.* Committee on Science and Technology. 2007. 110 Cong., 1st sess. SN 110-34, June 5. http://purl.access.gpo.gov/GPO/LPS88842.

Williams, Becky. 2009. Thousands of World of Warcraft fans descend on southern California for Blizzard's epic gaming convention. *Telegraph.co.uk.* August 24. http://tinyurl.com/ref-williams-2009.

Wray, Richard. 2009. Digital economy bill promises action on piracy, games and ITV regional news. *The Guardian*, November 18. http://www.guardian.co.uk/media/2009/nov/18/digital-economy-bill.

Zentner, Alejandro. 2006. Measuring the Effect of File Sharing on Music Purchases. *The Journal of Law and Economics* 49, no. 1: 63-90. doi:doi:10.1086/501082.